Analyse bayésienne robuste de lois d'extrêmes

Pablo Soto, Tiziano Fassina

March 2025

Let

$$F(x,\theta) = \exp\left\{ -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\},$$

be the cdf of the generalized extreme value law, whose density is

$$f(x;\theta) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi-1} \exp\left\{ -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\}.$$

We consider θ as the vector of $\theta = (\mu, \sigma, \xi)$.

Given some observed data $\mathbf{x} = (X_1, ..., X_m)$, their likelihood is given by $l_{\mathbf{x}}(\theta) = \prod_{j=1}^{m} f(X_j;\theta)$.

We want to solve, for a grid of values of x, the following optimization problem: Given $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_{n+1})$ and $\mathbf{p} = (p_1, ..., p_{n+1}) \in [0,1]^{n+1}$, find the supremum and infimum of

$$G_x(\boldsymbol{\theta}, \mathbf{p}) = \frac{\sum_{j=1}^{n+1} F(x;\theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j},$$

with the constraints:

- $\sum_{j=1}^{n+1} F(x_{q_i}) p_j = q_i, i = 1, ..., n$
- $\sum_{j=1}^{n+1} p_j = 1$

We will now prove some results that enable us to guarantee that $G_x(\boldsymbol{\theta}, \mathbf{p})$ takes a range of possible values, giving us a lower bound for the supremum and a higher bound for the infimum. In order to do that we will create some limit scenarios:

Lemma 1. *Let $q \in [0,1]$ and $x_0 \in \mathbb{R}$. Then, there exists a sequence $\theta_n = (\mu_n, \sigma_n, \xi_n)$ such that $\forall x \in \mathbb{R}, \lim_{n\to\infty} F(x;\theta_n) = q\mathbb{1}_{(x_0,\infty)}(x)$.*

Proof. Let $\xi_n = n, \mu_n = x_0 + (-\log q)^n$ and $\sigma_n = n(-\log q)^n$. Then,

$$\begin{aligned}
\lim_{n\to\infty} F(x, \theta_n) &= \lim_{n\to\infty} \exp\left\{-\left[1 + n\left(\frac{x - x_0 - (-\log q)^n}{n(-\log q)^n}\right)\right]_+^{-1/n}\right\} \\
&= \lim_{n\to\infty} \exp\left\{-\left[x - x_0\right]_+^{-1/n}(-\log q)\right\} \\
&= \exp\left\{-\lim_{n\to\infty}\left[x - x_0\right]_+^{-1/n}(-\log q)\right\} \\
&= q\mathbb{1}_{(x_0,\infty)}(x),
\end{aligned}$$

where I used the convention that $0^{-x} = \infty$ for some $x > 0$ as it is the only one that enables $F(x; \theta)$ to be a cdf ($\lim_{\epsilon\to 0_+} \epsilon^{-x} = \infty$ for $x > 0$, which means that $\lim_{\epsilon\to 0_+} F(x_0 + \epsilon, \theta_n) = 0$ and $F(x; \theta_n)$ must be increasing in x). □

Lemma 2. *Let $q \in [0, 1]$ and $x_0 \in \mathbb{R}$. Then, there exists a sequence $\theta_n = (\mu_n, \sigma_n, \xi_n)$ such that $\forall x \in \mathbb{R}, \lim_{n\to\infty} F(x; \theta_n) = \mathbb{1}_{[x_0,\infty)}(x) + q\mathbb{1}_{(-\infty,x_0)}(x)$.*

Proof. Let $\xi_n = -n, \mu_n = x_0 - (-\log q)^{-n}$ and $\sigma_n = n(-\log q)^{-n}$. Then,

$$\begin{aligned}
\lim_{n\to\infty} F(x, \theta_n) &= \lim_{n\to\infty} \exp\left\{-\left[1 - n\left(\frac{x - x_0 + (-\log q)^{-n}}{n(-\log q)^{-n}}\right)\right]_+^{1/n}\right\} \\
&= \lim_{n\to\infty} \exp\left\{-\left[x_0 - x\right]_+^{1/n}(-\log q)\right\} \\
&= \exp\left\{-\lim_{n\to\infty}\left[x_0 - x\right]_+^{1/n}(-\log q)\right\} \\
&= \mathbb{1}_{[x_0,\infty)}(x) + q\mathbb{1}_{(-\infty,x_0)}(x),
\end{aligned}$$

□

Theorem 1. *Let $l \in [[1, n]]$ and $x_0 \in [x_{q_l}, x_{q_{l+1}})$. Suppose there exists a $q \in [q_l, q_{l+1}]$ such that $q(q_1 + q_n) < q^2 + q_1$. Then, there exists a value of $(\boldsymbol{\theta}, \boldsymbol{p}) = (\theta_1, ..., \theta_{n+1}, p_1, ..., p_n)$ that follows the constraints such that*

$$G(x_0; \boldsymbol{\theta}, \boldsymbol{p}) \in (q - \epsilon, q + \epsilon), \quad \forall \epsilon > 0.$$

This is our main result. It enables us to guarantee that there exist a value of $(\boldsymbol{\theta}, \mathbf{p})$ such that $G_x(\boldsymbol{\theta}, \mathbf{p})$ is as close as we want from the value q. If we have that for all q_j, $q_j(q_1 + q_n) < q_j^2 + q_1$, then we have a higher bound if the infimum:

$$\inf G_x(\boldsymbol{\theta}, \mathbf{p}) \leq \sum_{j=1}^{n} q_j \mathbb{1}_{(x_{q_j}, x_{q_{j+1}})}(x),$$

and a lower bound of the supremum:

$$\sup G_x(\boldsymbol{\theta}, \mathbf{p}) \geq \sum_{j=0}^{n} q_{j+1} \mathbb{1}_{(x_{q_j}, x_{q_{j+1}})}(x).$$

Proof. The idea is to have $\forall j \in [[1,n]]$, $F(x_0,\theta_j) \in (q-\epsilon, q+\epsilon)$. In this way,

$$G(x_0,\boldsymbol{\theta},\mathbf{p}) = \frac{\sum_{j=1}^{n+1} F(x_0;\theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j} \leq \frac{(q+\epsilon)\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j} = q+\epsilon.$$

We get $G(x_0,\boldsymbol{\theta},\mathbf{p}) \geq q-\epsilon$ in an analogous way.

In order to do this, we will make use of the functions of the previous lemmas. We first work using the functions in the limit to then bound the error of the sequences. We define for $j \in [[2, l+1]]$ the function $F_j(x) = q\mathbb{1}_{[x_{q_{l-j+2}},\infty)}(x)$, which corresponds to the limit of Lemma 1.2 using $x_0 = x_{q_{l-j+2}}$; for $j \in [[l+2, n+1]]$ the function $F_j(x) = \mathbb{1}_{[x_{q_{n+l-j+2}},\infty)}(x) + q\mathbb{1}_{(-\infty, x_{q_{n+l-j+2}})}(x)$, which corresponds to the limit of Lemma 1.3 using $x_0 = x_{q_{n+l-j+2}}$; and for $j=1$, $F_1(x) = q\mathbb{1}_{[x_0,\infty)}(x)$. Note that for all functions, $F_j(x_0) = q$.

We define the vectors $\mathbf{x} = (x_{q_1}, x_{q_2}, ..., x_{q_l}, x_0, x_{q_{l+1}}, ..., x_{q_n}) \in \mathbb{R}^{n+1}$, $\mathbf{q} = (q_1, ..., q_n, 1) \in \mathbb{R}^{n+1}$ and $\mathbf{p} = (p_1, ..., p_{n+1}) \in [0,1]^{n+1}$. We then build the matrix $F = (F_{ij})_{i,j=1}^{n,n+1} = (F_j(x_i))_{i,j=1}^{n,n+1}$. The entries of the matrix are given by the formula

$$F_{ij} = q\mathbb{1}_{[[l+2-j, n+l+1-j]]}(i) + q\mathbb{1}_{[[n+l+2-j,n]]}(i).$$

Then, finding a "good" $\mathbf{p}$ so that the constraints are fulfilled is equivalent of finding a non-negative solution of the linear system

$$\bar{F}\mathbf{p} = \mathbf{q},$$

where $\bar{F}$ is the matrix F with a row of ones appended below. We can manually find a solution of the system. Subtracting consecutive equations, we get that for $j \in [[1, l-1]]$, $p_{l-j+1} = \frac{q_{j+1}-q_j}{q}$ and $p_1 = \frac{q-q_l}{q}$. Using the rest of the equations, we get that for $j \in [[1, n-l-1]]$, $p_{l+j+1} = \frac{q_{n-j+1}-q_{n-j}}{1-q}$ and $p_{n+1} = \frac{q_{l+1}-q}{q}$. All these values are in $[0,1]$ as the q_i are increasing and less than 1. Then, we get that

$$1-p_{l+1} = \frac{q-q_l}{q} + \sum_{j=1}^{l-1} \frac{q_{j+1}-q_j}{q} + \sum_{j=1}^{n-l-1} \frac{q_{n-j+1}-q_{n-j}}{1-q} + \frac{q_{l+1}-q}{q} = \frac{q-q_1}{q} + \frac{q_n-q}{1-q}.$$

We need to check that the value is smaller than 1, which is equivalent to $q(q_1+q_n) \leq q^2 + q_1$. So, if this condition holds, we have a "limit solution". If we suppose that q is in the interior of the set (i.e. $q_l < q < q_{l+1}$ and $q(q_1+q_n) < q^2+q_1$), then the vector $\mathbf{p}$ is in the interior of the probability simplex of dimension $n+1$. As $\mathbf{p}$ is the solution of a linear system, this means that there exists an $\epsilon_1 > 0$ such that for all perturbations $\tilde{F}$ of $\bar{F}$ such that $\|\bar{F} - \tilde{F}\| \leq \epsilon_1$, then $\tilde{F}$ is invertible and the solution $\mathbf{p'} = \tilde{F}^{-1}\mathbf{q}$ is in the probability simplex. Furthermore, there also exists an $\epsilon_2 > 0$ such that if $\forall i, j \in [[1, n+1]]$, $|\tilde{F}_{ij} - \bar{F}_{ij}| < \epsilon_2$, then $\|\bar{F} - \tilde{F}\| \leq \epsilon_1$.

Lemmas 1.2 and 1.3 show that for all $j \in [[1, n+1]]$, there exist sequences $\theta_{j,k}$ such that $F(x; \theta_{j,k})$ converges punctuality to $F_j(x)$. Then, we can pick a k_0 such that $|F(x_j, \theta_{j,k_0}) - F_j(x_j)| < \min(\epsilon_2, \epsilon)$, for all elements x_j of the vector $\mathbf{x}$. Then, we have the condition we searched at the beginning of the proof and some probability vector $\mathbf{p}^{k_0}$ such that $(\boldsymbol{\theta}^{k_0}, \mathbf{p}^{k_0})$ fulfills the constraints.

□

These are all "data-free" results, in the sense that they don't use the values of the likelihood. This is interesting as it enables us to have a lower bound for the range of G_x even without taking measurements. If we want to have better results the only idea I have is to use this likelihood, but I think it will make calculations much more difficult.